NEXUS Ingestion

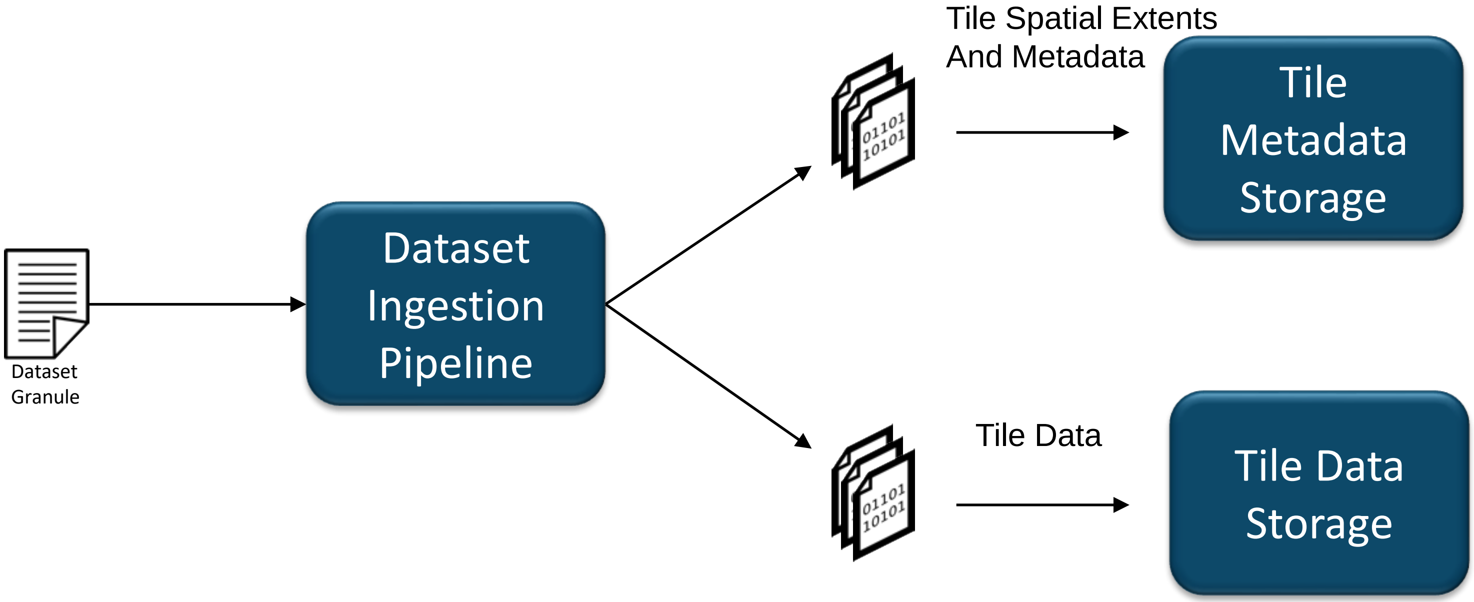
Frank Greguska

Jet Propulsion Laboratory, California Institute of Technology

Ingestion Pipeline

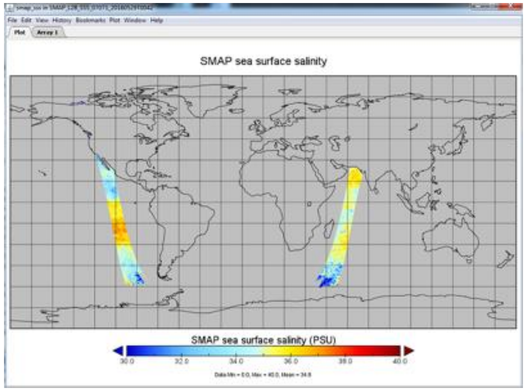
- What is a Tile?
  - A collection of nd-arrays containing measurement data and its associated metadata
  - One granule becomes multiple tiles
  - Allows for fast spatial lookup of array data

- Horizontally Scalable Storage
  - Apache Solr Cloud
  - Apache Cassandra, Amazon S3



Ingestion Pipeline

- Ingestion pipeline supports multiple tiling algorithms
  - L2 Swath Data
  - L3/L4 Gridded Data


SMAP sea surface salinity
File Edit View History Bookmarks Plot Window Help
Plot
Array 1
SMAP sea surface salinity (PSU)
30.0
32.0
34.0
36.0
38.0
40.0

JPL/CAP L2B SMAP Sea Surface Salinity

L3/L4 Grid Tiling Algorithm:

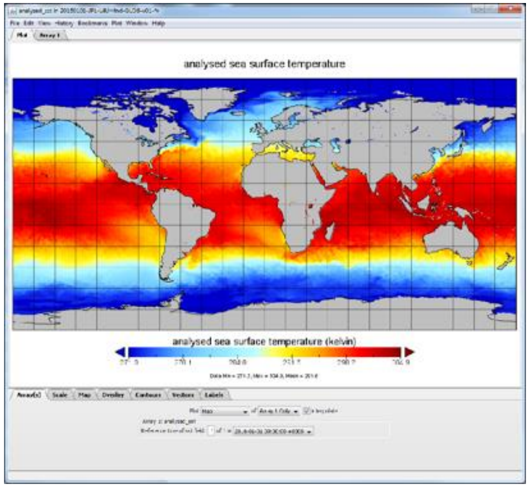
$c = Number\ of\ tiles\ desired$

$d = Number\ of\ dimensions$

$L_d = Length\ of\ dimension\ d$

$S_d = Step\ size\ for\ dimension\ d$

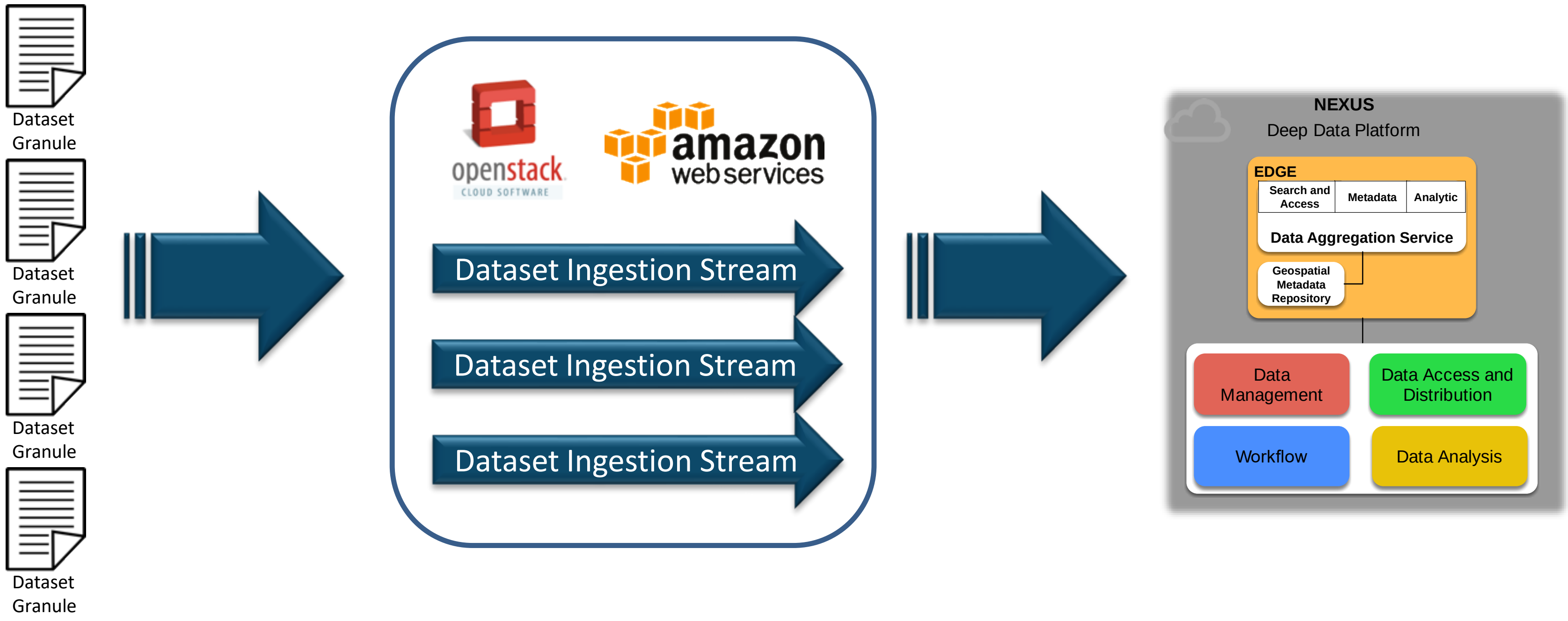
$$S_d = \left\lfloor \frac{L_d}{\sqrt[d]{c}} + \frac{1}{2} \right\rfloor$$


analysed sea surface temperature
Plot
Array 1
analysed sea surface temperature (kelvin)
Array(s)
Scale
Map
Overlay
Contours
Vectors
Labels

MUR-JPL-L4-GLOB-v4.1 Analyzed Sea Surface Temperature

Ingestion Pipeline

- Pipelines can run in parallel
- Individual pipeline modules can be scaled horizontally
- Pipelines deployable to the cloud



Ingestion Pipeline

- Pluggable validation checks

```
def filter_empty_tiles(self, tile):
    # Only supply data if there is actual values in the tile
    if tile.data.size - numpy.count_nonzero(numpy.isnan(tile.data)) > 0:
        yield tile.data
    else:
        print "Discarding data %s from %s because it is empty" % (tile.section_spec, tile.granule)
```

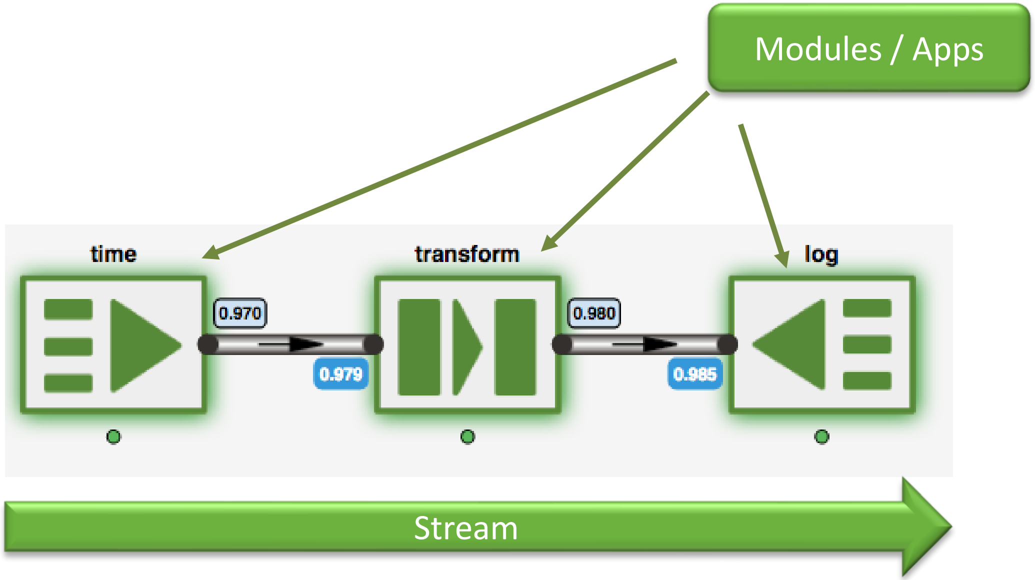
- Data transformation

```
def transform(self, tile):

    tile.data.longitudes[longitudes > 180] -= 360

    yield tile.data
```

Ingestion Pipeline

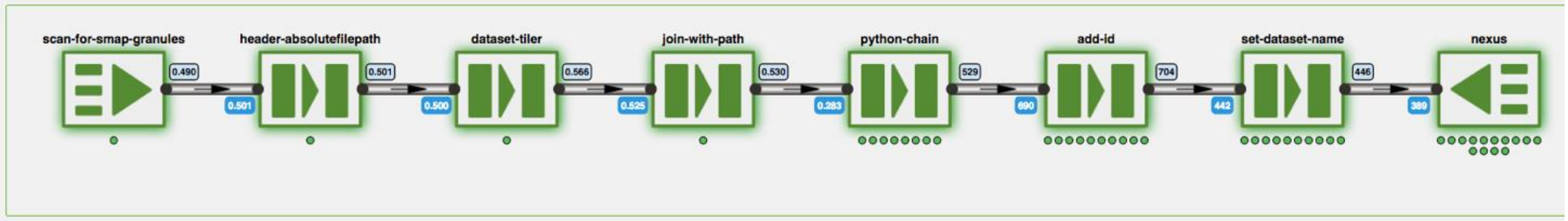
- Spring XD
  - http://projects.spring.io/spring-xd/
  - Current production release
  - Additional software components: Zookeeper, Kafka, Redis

- Spring Cloud Data Flow
  - http://cloud.spring.io/spring-cloud-dataflow/
  - Redesign of Spring XD



Ingestion Pipeline

- Current Deployments
  - Bare Metal NASA AIST-funded Deep Data Computing Environment (DDCE) at JPL
  - Mirantis OpenStack at JPL
  - NASA AIST Managed Cloud Environment (AMCE)
- Applications are connected to form ingestion streams
- Configurable to handle different datasets
- Scalable across compute resources
- Resilient to failure



Stream for JPL/CAP L2B SMAP Sea Surface Salinity